BOTNET DETECTION IN IoT

Synopsis of the PhD Thesis

Submitted By

Deepa Krishnan

Roll. No. 70821018010

Department of Computer Engineering

Supervisor

Dr. Pravin Shrinath

Submitted in partial fulfilment of the requirements of
DOCTOR OF PHILOSOPHY (Engineering)

MUKESH PATEL SCHOOL OF TECHNOLOGY MANAGEMENT & ENGINEERING

SVKM's NARSEE MONJEE INSTITUTE OF MANAGEMENT STUDIES (Declared as Deemed-to-be University Under Section 3 of the UGC Act, 1956)

V. L. Mehta Road, Vile Parle (West)

MUMBAI –400056

August 2024

Abstract

The rapid expansion of the Internet of Things (IoT) has transformed various fields, including healthcare, industrial automation, and smart cities, by enabling seamless data exchange and connectivity through interconnected devices. However, this growth has also introduced substantial cybersecurity challenges, most notably the rise of IoT botnets. These botnets, which consist of compromised devices controlled by malicious actors, exploit the inherent vulnerabilities within IoT environments to carry out large-scale attacks, jeopardizing data integrity, service availability, and overall system security. This research aims to address the urgent need for robust and efficient botnet detection mechanisms that are specifically designed for the resource-constrained and heterogeneous nature of IoT devices.

The central focus of this study is the development of advanced methodologies to enhance botnet detection while tackling issues such as class imbalance in security datasets and ensuring resilience against adversarial attacks. A significant contribution of this research is the introduction of a multi-stage feature selection technique that combines joint mutual information with maximum relevance and minimum redundancy scores. This approach substantially improves classification accuracy, particularly in imbalanced datasets.

Moreover, the research proposes an algorithmic weighted method to further enhance classification performance in highly imbalanced datasets, without incurring the computational costs typically associated with traditional re-sampling techniques. The creation of a stacked Multilayer Perceptron (MLP) classifier, featuring class-wise feature selection and adaptive reconstruction error thresholding, demonstrates remarkable effectiveness in detecting both known and novel attacks, thereby enhancing the reliability of intrusion detection systems.

Additionally, this research introduces a denoising autoencoder-based approach that is resilient against both conventional and advanced adaptive adversarial attacks. This method ensures high classification accuracy and reduces the rate of misclassification under various adversarial conditions, providing a lightweight and efficient defense mechanism well-suited for IoT devices with limited resources.

The practical implications of these findings are profound, offering improved security for IoT networks, reduced computational demands, enhanced incident response capabilities, and robust defenses against adversarial threats. These contributions represent a significant advancement in the field of IoT botnet detection, ensuring resilient and adaptive security solutions that are essential for the safe and reliable deployment of IoT technologies across diverse sectors. Future research should explore more sophisticated adversarial attack

strategies and develop advanced defense mechanisms by leveraging progress in adversarial machine learning. Additionally, there is a need to focus on improving online learning algorithms to adapt to concept drift in real-time, ensuring that detection models remain accurate and resilient as data distributions change over time.

Contents

1 Introduction 5
1.1 Background of the Study 5
1.2 Need for Botnet Detection in IoT 6
1.3 Significance of the Proposed Research 7
1.4 Scope of the Proposed Research 8
1.5 Summary 8
2 Research Objectives 9
3 Literature Review 10
3.1 Recent works in detection of botnet attacks in IoT 10
3.1.1 Research Gaps Identified 11
3.2 Recent works in detection of unseen attacks attacks in IoT 11
3.2.1 Research Gaps Identified 12
3.3 Recent works in Adversarial attack detection 13
3.3.1 Research Gaps Identified 14
3.4 Summary of Research Gaps 14
4 Contributions 15
4.1 Design of Multi-stage Feature Selection and Weighted Approach . 16
4.1.1 Proposed Multi-stage Feature Selection 16
4.1.2 Algorithmic Weighted Approach for Class Weight Generation 17
4.1.3 Results and Analysis 17
4.2 Proposed Scheme for Detection of Known and Unknown Attacks . 18
4.2.1 Proposed Scheme for Detection of Class-wise Feature Selection 18
4.2.2 Design of Stacked Multi-Classifier and Detection of Known Attacks 21
4.2.3 Proposed Scheme for Detection of Unknown Attacks . . 21
4.2.4 Result and Analysis 22
4.3 Design of Robust Botnet Detection Model against Adversarial Attacks 27
4.3.1 Proposed Defense using Shallow Denosing AutoEncoder 27

4.3.2 Result and Analysis 30
5 Conclusion 33
6 Future Scope 34
7 Publications 35

1 Introduction

1.1 Background of the Study

The Internet of Things (IoT) has fundamentally transformed computing and connectivity, merging wireless technologies, embedded systems, and ubiquitous computing to create a network of interconnected physical devices. These devices, ranging from everyday objects to industrial machinery, autonomously collect and exchange data. The rapid growth of IoT has been fueled by advancements in miniaturization, cost-effective sensors, and widespread high-speed internet connectivity, revolutionizing various sectors such as healthcare, transportation, smart cities, agriculture, and industrial automation [1]. In healthcare, IoT devices like wearable sensors and remote monitoring systems enable continuous health tracking and personalized treatment plans. Industrial IoT applications optimize operational efficiency through predictive maintenance and real-time monitoring, while agriculture benefits from smart sensors for soil monitoring and irrigation automation. Smart cities leverage IoT to improve public services and resource allocation. However, the proliferation of IoT devices has also led to increased cybersecurity challenges, with the sheer volume and diversity of connected devices expanding the attack surface for malicious actors [2].

One of the significant threats facing IoT environments is the emergence of botnets, which are networks of compromised devices controlled by a central command. These botnets exploit the resource constraints and diverse communication protocols of IoT devices to orchestrate large-scale attacks, including Distributed Denial of Service (DDoS) assaults [3]. In recent years, notable instances of IoT botnet attacks, such as the Mirai botnet, have underscored the urgency of addressing this security threat. The sheer scale and impact of these attacks have raised awareness about the vulnerabilities inherent in IoT deployments [4]. The most recent Nokia Threat Intelligence Report reveals a substantial surge in IoT botnet Distributed Denial of Service (DDoS) attacks over the past year, increasing fivefold. This rise in malicious activity is attributed to profit-driven hacking collectives, particularly observed in the aftermath of Russia's invasion of Ukraine. The escalation of IoT botnet DDoS attacks, targeting insecure devices, has spread globally, impacting telecom networks and critical infrastructure. The number of participating IoT devices has surged from 200,000 to over 1 million, constituting over 40% of all DDoS traffic.

The predominant malware affecting telecommunication networks involves a bot malware employing scanning tactics for vulnerable devices, a strategy associated with various IoT botnets. The widespread deployment of IoT devices globally, ranging from smart appliances

to medical sensors, contributes to the vulnerability, given the lax security measures in place. Additionally, the report highlights a concerning doubling in trojans targeting personal banking information on mobile devices, posing a heightened risk to millions of users globally. The report underscores the urgent need for robust and elevated security measures within the realm of IoT. As a result, researchers and cybersecurity experts are increasingly focusing on developing strategies to detect and mitigate the presence of botnets within IoT networks. This research aims to contribute to this evolving field by exploring innovative approaches for the effective detection of botnets in IoT environments, considering the unique challenges posed by the heterogeneity, resource constraints, and dynamic nature of IoT

1.2 Need for Botnet Detection in IoT

The imperative for botnet detection in IoT arises from the unique vulnerabilities inherent in the expansive network of interconnected devices. Traditional cybersecurity measures, developed for more conventional computing environments, often fall short in adapting to the intricacies of the IoT paradigm. The heterogeneity of IoT devices, coupled with their resource constraints, creates a complex and dynamic environment that demands specialized detection techniques. The consequences of a successful botnet attack on IoT can be profound, leading to service disruptions, unauthorized access, and compromise of sensitive information. Thus, the need to develop effective and adaptive botnet detection mechanisms tailored to the specific characteristics of IoT ecosystems is critical for safeguarding the burgeoning IoT landscape [5].

One of the primary challenges in IoT botnet detection lies in the sheer diversity of devices and communication protocols. The wide spectrum of IoT devices, ranging from low-power sensors to high-performance industrial machines, introduces a multitude of entry points for potential compromise. Furthermore, the heterogeneity extends to communication protocols, making it challenging to establish universal detection methods. Additionally, the resource limitations of many IoT devices necessitate the development of lightweight and efficient detection mechanisms that do not unduly burden the devices themselves [6].

The potential impact of botnet-driven attacks in the IoT realm extends beyond individual devices to encompass critical infrastructure and services. With IoT devices increasingly embedded in mission-critical systems such as healthcare monitoring, energy grids, and transportation networks, the repercussions of a botnet attack can extend to public safety and national security [7]. Recognizing this, the research on botnet detection in IoT seeks to address these challenges by developing innovative and context-aware detection techniques. By enhancing the security posture of IoT ecosystems, this research

aims to contribute to the continued growth and integration of IoT technologies across various domains while mitigating the inherent risks associated with botnet threats.

1.3 Significance of the Proposed Research

The research outlined is of significant importance due to its potential to address critical challenges in botnet detection within the Internet of Things (IoT) environment. As the IoT continues to expand and connect more devices, the risk of botnet attacks becomes more pronounced. Botnets can exploit vulnerabilities in IoT devices to carry out malicious activities such as distributed denial-of-service (DDoS) attacks, unauthorized data access, and device hijacking. Therefore, developing effective botnet detection mechanisms tailored to the IoT ecosystem is crucial for ensuring the security and integrity of IoT deployments.

One key significance of this research is its focus on developing computationally efficient feature selection techniques. IoT devices often have limited computational resources, making it challenging to implement complex detection models. By designing feature selection methods that are efficient yet effective, this research can enable the development of lightweight botnet detection solutions that can run on resource-constrained IoT devices. This would significantly enhance the overall security of IoT ecosystems by enabling real-time detection and mitigation of botnet threats. Another significant aspect of this research is its emphasis on addressing class imbalance in botnet detection without relying on traditional resampling methods. Class imbalance is a common issue in security datasets, where one class normal behavior significantly outnumbers another class e.g., botnet activity. By formulating innovative approaches to mitigate class imbalance challenges, this research can improve the reliability and accuracy of botnet detection models, leading to more robust security solutions for IoT environments.

Additionally, the development of a class-wise feature selection strategy for known attack detection is highly significant. This strategy can improve the precision and sensitivity of known attack detection mechanisms by prioritizing features relevant to each specific attack class. By tailoring feature selection to individual attack types, this research can enhance the overall efficacy of botnet detection in IoT environments, enabling more targeted and effective mitigation of known attacks. Furthermore, the research's focus on techniques to detect unseen attacks in IoT environments is crucial for ensuring the resilience of security solutions. As IoT threat landscapes evolve, the ability to detect emerging threats without prior knowledge is essential. By devising methodologies for the detection of previously unseen or novel attacks, this research can enhance the overall security posture of IoT ecosystems, safeguarding them against evolving threats.

1.4 Scope of the Proposed Research

The scope of this research encompasses the development and evaluation of advanced methodologies for botnet detection in IoT environments, with a focus on computational efficiency and robustness against evolving threats. The research aims to design novel feature selection techniques that are specifically tailored to the limited computational capabilities of IoT devices. These techniques will prioritize the most relevant features for botnet detection, ensuring that the detection models can operate efficiently without compromising on accuracy. Additionally, the research addresses the challenge of class imbalance in security attack data, proposing innovative methods to mitigate this issue without relying on traditional resampling techniques. This aspect of the research is critical for enhancing the reliability of detection models in real-world scenarios where class imbalance is prevalent.

Furthermore, the research extends to the detection of previously unseen attacks and the evaluation of adversarial robustness of botnet detection models. By developing methodologies that can identify novel threats and withstand adversarial attacks, the research aims to create resilient security solutions for IoT networks. This includes the creation of adversarial attack samples, the design of defense mechanisms, and thorough evaluation of the models' performance under adversarial conditions. The comprehensive approach ensures that the proposed solutions are not only effective in current threat landscapes but also adaptable to future challenges. Overall, the research scope is geared towards establishing a robust, efficient, and adaptive framework for IoT botnet detection, addressing critical challenges in feature selection, class imbalance, and adversarial robustness.

1.5 Summary

In the subsequent Sections, we delved into a comprehensive analysis of existing botnet detection techniques, aiming to identify their strengths, limitations, and applicability within IoT environments. We will explore innovative approaches for developing computationally efficient feature selection techniques, addressing class imbalance, and implementing a class-wise feature selection strategy for known attack detection. Additionally, we have investigated techniques to detect previously unseen attacks in IoT environments, enhancing the overall robustness of our detection framework. Furthermore, we have evaluated the adversarial robustness of the developed botnet detection framework under various attack scenarios, providing insights into its resilience against sophisticated adversaries. Through these efforts, we aim to contribute to the advancement of botnet detection mechanisms tailored to IoT ecosystems, enhancing the security posture of IoT deployments

2 Research Objectives

The research objectives outlined in this study focus on addressing critical challenges in botnet detection within the Internet of Things (IoT) environment.Each objective is designed to contribute to the development of innovative and effective security solutions tailored to the unique characteristics and complexities of IoT ecosystems.

- **Design of Computationally Efficient Feature Selection Techniques:**
  - **RO1**: Develop novel methodologies for selecting relevant features in IoT environments, emphasizing computational efficiency to accommodate the resource constraints of IoT devices
  - **Rationale:** Efficient feature selection is crucial for optimizing the performance of botnet detection models on devices with limited computational capabilities.
- **Design of Techniques for Addressing Class Imbalance in Botnet Detection:**
  - **RO2**:Formulate innovative approaches to mitigate class imbalance challenges in botnet detection, without relying on traditional resampling methods.
  - **Rationale**:Class imbalance is a common issue in security datasets; designing techniques to address this imbalance ensures more reliable and accurate botnet detection models..
- **Class-wise Feature Selection Strategy for Known Attack Detection**
  - **RO3**: Develop a class-wise feature selection strategy for enhancing the detection of known attacks by prioritizing features relevant to each specific attack class.
  - **Rationale**:Tailoring feature selection to individual attack types enhances the precision and sensitivity of detection mechanisms, improving the overall efficacy of known attack detection.
- **Techniques to Detect Unseen Attacks in IoT**
  - **RO4**:Devise methodologies for the detection of previously unseen or novel attacks within IoT environments, focusing on adaptive and learning-based approaches..
  - **Rationale**:As IoT threat landscapes evolve, the ability to detect emerging threats without prior knowledge is essential for ensuring the resilience of security solutions.
- **Evaluation of Adversarial Robustness of Developed Security Solutions**

- **RO5**:Conduct a comprehensive evaluation of the developed security solutions, specifically assessing their resilience against intentional attacks aimed at evading detection.
- **Rationale**:Adversarial robustness testing ensures that the proposed security solutions can withstand deliberate attempts to deceive or circumvent detection mechanisms..

3 Literature Review

The following section provides a comprehensive review of recent research in the field of security attack detection in IoT networks, with a focus on the detection of botnet attacks, unseen attacks, and adversarial attacks. The review highlights various methodologies, techniques, and machine learning algorithms used in these areas, discussing their strengths, limitations, and research gaps

3.1 Recent works in detection of botnet attacks in IoT

Recent studies have proposed several innovative approaches to address the challenge of imbalanced data in IoT security attack detection. One such approach, mixture localization-based outliers detection, combines Gaussian Mixture Modelling [8] and the local outlier function (LOF) to detect outliers in network data, achieving high detection rates but lower rates for minority classes. Traditional machine learning algorithms, including C4.5 Decision Tree, Random Forest, SVM, and Naïve Bayes, have been employed with promising results for majority classes, although lacking detailed information on model building time and performance scores for minority classes. To tackle class imbalance, techniques such as the SMOTE-ENN re-sampling technique have been utilized, showing satisfactory performance with challenges encountered in certain datasets [9].

Researchers have also explored the use of Principal Component Analysis (PCA) for feature reduction and proposed a lightweight neural network called Lock Edge for optimizing machine learning models in IoT networks. While these approaches demonstrated high detection rates for certain attacks, they showed limitations in detecting other attacks due to their simplicity. To address imbalanced data distribution, advanced techniques like EG-SMOTE combined with Growing Self Organization Map (GSOM) and hyperdimensional computing have been introduced, enhancing energy efficiency but potentially leading to overfitting [10]. Privacy-preserving techniques and feature transformation methods have also been employed, showing improved AUC values but suboptimal performance for certain attacks [11].

Collaborative attack detection has been explored through approaches like the voting classifier with feature reduction, which achieved better classification performance despite limitations in sample size [12]. Early-stage detection of IoT-based botnets has been another focus, utilizing machine learning algorithms for traffic classification and individual bot identification. However, some approaches have yielded low detection rates and F1 Scores for minority classes. Additionally, studies on federated learning in IoT networks have revealed challenges in achieving high accuracy, precision, and F1 Scores, particularly in cases of skewed distributions, highlighting the need for further research in this area. Overall, these studies emphasize the complexities and challenges associated with imbalanced learning in IoT security attack detection, underscoring the importance of developing more efficient and effective solutions.

3.1.1 Research Gaps Identified

In the literature surveyed, wrapper-based feature selection techniques are commonly utilized for selecting relevant features. While effective, these methods are computationally expensive, making them inefficient for IoT devices. Therefore, there is a need for less computationally intensive approaches that can select the most relevant and minimally redundant set of features.Resampling-based approaches are frequently employed to handle imbalanced classes. Techniques like SMOTE [13], [14], EG-SMOTE [10], G-SMOTE [10], and WCGAN [15] are used to address class imbalance by either under-sampling majority classes or oversampling minority classes. However, these approaches can lead to the loss of valuable data or the addition of spurious data, increasing training and prediction times and adding overhead for attack detection in IoT.

Despite the use of over-sampling techniques, many approaches in the literature failed to improve performance scores for detecting classes with very few samples. Furthermore, none of the reviewed researchers have analyzed the energy consumption of their implemented machine learning security models or assessed associated overheads such as training time, prediction time, CPU utilization, and memory utilization.While hybrid deep-learning models and ensemble models show impressive performance scores, they often come with significantly higher training times and parameter requirements.

3.2 Recent works in detection of unseen attacks attacks in IoT

The literature provides a thorough exploration of techniques for detecting previously unseen attacks. In a study conducted by [16], the researchers introduce an innovative method aimed at identifying such attacks. Their work involves assessing the performance of trained models in classifying new samples from the same type of attack, but originating

from a different dataset. Although the integration of trained models with closely related novel data was anticipated to preserve high classification accuracy, the results on the unseen dataset indicate a decline in accuracy, precision, and recall. Importantly, the model's performance remains robust only when applied to the dataset on which it was initially trained.

An alternative strategy for detecting unseen attacks involves the use of ensemble deep learning architectures, as explored by [17]. This study evaluates the model's classification accuracy when it correctly identifies an attack test case as belonging to one of the attack categories, signaling that the model recognized the instance as an attack but was uncertain about its precise classification. If an attack test instance is incorrectly classified as benign, it is considered a classification error. Although the study demonstrated accurate classification in some instances, certain proposed methods showed near-zero detection accuracy for particular attack classes.

In a similar vein, [18] explores the detection of unseen attacks using a Correntropy-based One-class ensemble model with a Gaussian mixture. The study emphasizes deploying this model on edge nodes, with experimentation carried out on the NSL-KDD and UNSW-NB15 datasets. Additionally, [19] introduces a transfer learning-based method for identifying zero-day attacks in IoT environments. This technique involves using the BoT-IoT dataset as the source domain to acquire knowledge, which is then applied to the UNSW-NB15 dataset as the target domain. Moreover, [20] presents an innovative two-stage approach that combines deep neural networks with transfer learning to accurately detect zero-day malware by analyzing hardware events via images. Despite the advantages, a key limitation of the transfer learning method is its dependence on pre-trained models from a different dataset, which may inadvertently capture features unique to the source domain, leading to poor generalization on unseen data.

In a systematic literature review by [21], diverse methodologies, techniques, and machine learning and deep learning algorithms used for the detection of zero-day attacks are analyzed. Despite technological advancements and the robust processing capabilities of machine learning and deep learning algorithms, the identification of entirely new or unknown attacks remains an ongoing research frontier.

3.2.1 Research Gaps Identified

Recent literature on identifying previously unseen attacks has revealed several key gaps and limitations in existing methodologies. In the study by [16], while evaluating trained models, a decrease in performance metrics was observed when applied to unseen datasets, indicating a challenge in generalizing to novel attack scenarios. Similarly, the study by [17] achieved accurate classification for certain cases but exhibited detection accuracy close to

zero for specific classes, suggesting limitations in detecting certain types of unseen attacks effectively. Additionally, the work by [18], which focused on edge node deployment, was limited to specific datasets (NSL-KDD and UNSW-NB15), raising questions about the model's performance in other scenarios.

Furthermore, the transfer learning approach used by [19] relied on pre-trained models from a different dataset, potentially capturing source domain characteristics and limiting generalizability to unseen data in different domains. The two-stage methodology proposed by [20] for zero-day malware detection lacked evaluation with different datasets or in real-world scenarios, leaving a gap in understanding its practical effectiveness. Lastly, despite the systematic literature review conducted by [21] to identify emerging research trends and challenges, there was a lack of discussion on potential solutions or future directions, limiting actionable insights for researchers in the field

3.3 Recent works in Adversarial attack detection

Recent literature on identifying previously unseen attacks in the field of security attack detection has revealed several key gaps and limitations in existing methodologies. Studies have shown a decrease in performance metrics when trained models are applied to unseen datasets, indicating challenges in generalizing to novel attack scenarios. Some approaches have achieved accurate classification for certain cases but exhibited low detection accuracy for specific classes, suggesting limitations in detecting certain types of unseen attacks effectively. Additionally, certain studies focusing on specific datasets raise questions about the models' performance in other scenarios, highlighting the need for more robust and generalized approaches [22].

In addressing adversarial attacks, which pose a significant challenge to the reliability of machine learning-based security systems, researchers have explored various strategies. Gradient-based attacks, such as the Fast Gradient Sign Method (FGSM), are commonly used to craft adversarial examples that deceive the model into making incorrect predictions. Defending against these attacks encompasses three primary strategic approaches: Architectural Modifications and Ensemble Mechanisms, Adaptations in the Training Process, and Modification and Pre-processing of input Data. These approaches aim to enhance a network's robustness by altering its architecture, customizing the loss function, and introducing noise or distortion to the input data to make it more difficult for attackers to craft effective adversarial examples [23].

Additionally, transfer learning approaches and hierarchical learning have been proposed to reduce vulnerabilities to adversarial attacks. Ensemble-based approaches, such as the AppCon [24] framework and the FGMD [25] framework, have been effective in detecting adversarial examples while maintaining performance in non-adversarial scenarios.

Adversarial training, where the network is exposed to adversarial examples during training, has also shown promise in enhancing robustness. However, these approaches may come with limitations, such as increased computational overhead and resource requirements, highlighting the ongoing challenges in developing effective defenses against adversarial attacks.

3.3.1 Research Gaps Identified

Several research gaps exist in the current landscape of defending neural networks against adversarial attacks. Firstly, while approaches like adversarial training and architectural modifications show promise in improving robustness against known attacks, they lack generalization to handle novel or unseen attack strategies effectively. Thus, there is a critical need for research to develop more generalized adversarial defense mechanisms. Secondly, the computational complexity and resource demands of adversarial training make it impractical for resource-constrained IoT environments, highlighting the necessity for more resource-efficient techniques. Additionally, existing research primarily focuses on identifying adversarial samples, potentially leading to a trade-off between false positives and false negatives, compromising the reliability of detection methodologies. Lastly, attackers can adapt their strategies in response to implemented defense mechanisms, yet none of the reviewed research concentrates on designing defenses capable of withstanding adaptive attackers and unfamiliar attack methodologies. Addressing these gaps is crucial to advancing the field of adversarial defense and ensuring the security of neural networks in practical applications.

3.4 Summary of Research Gaps

The literature review examines recent advancements in the detection of security attacks in IoT networks, focusing on botnet attacks, unseen attacks, and adversarial attacks. Researchers have proposed novel approaches such as mixture localization-based outliers detection, traditional machine learning algorithms, and ensemble methods to detect outliers and address class imbalance. However, challenges remain in generalization to novel attacks, resource efficiency, and trade-offs between false positives and false negatives. Additionally, approaches for detecting unseen attacks include ensemble deep learning architectures and transfer learning techniques, but they face challenges in maintaining performance on unseen datasets. Defending against adversarial attacks involves architectural modifications, adversarial training, and pre-processing of input data, but research gaps exist in generalization, resource efficiency, and adaptive attack defense. Addressing these gaps is crucial for enhancing the security and reliability of IoT networks against

evolving security threats.

4 Contributions

The contributions of the proposed research are as follows:

- **Novel Feature Selection Method:** Development of a new method for selecting features from datasets using joint mutual information score and maximum relevance and minimum redundancy score to identify the most relevant features.
- **Algorithmic Weighted Approach:** Introduction of a new algorithmic weighted approach to handle imbalanced datasets, ensuring effective training on all classes, including minority classes, without using traditional resampling techniques.
- **Modified Matrix Learning Vector Quantization:** Development of a modified matrix learning vector quantization approach to identify discriminative features for each target class, enabling the construction of a stacked MLP classifier for effective data classification.
- **Reconstruction Error Thresholding Technique:** Proposal of a technique for detecting unseen attacks using shallow auto-encoders to reduce model size and complexity. Analysis of the reconstruction error to detect anomalies or attacks not seen during training.
- **Adversarial Attack Generation and Evaluation:** Generation of adversarial attack samples using gradient-based attacks such as PGD and FGSM to evaluate the MLP model's robustness against adversarial perturbations and its generalization to unseen attacks.
- **Shallow Denoising Auto-encoder:** Proposal of a shallow Denoising Auto-encoder to restore the MLP model's classification ability under adversarial attacks, ensuring the model remains robust in the face of adversarial attacks.

These contributions aim to enhance the performance, robustness, and efficiency of machine learning models, especially in IoT security applications, where data complexity and susceptibility to attacks are significant challenges.

4.1 Design of Multi-stage Feature Selection and Weighted Approach

4.1.1 Proposed Multi-stage Feature Selection

Feature selection plays a crucial role in the performance of machine learning models, as it determines which features contribute most effectively to the model's learning ability. The goal is to select a subset of features that maximizes relevance to the target variable while minimizing redundancy among features. One common technique used for this purpose is the minimum redundancy maximum relevance mRMR method.In the proposed technique, we extend the mRMR method to capture non-linear interactions between variables. The basic idea behind mRMR is to select features that have high relevance with the target variable and low redundancy with previously selected features. The relevance of a feature is typically measured using the F-statistic score, which indicates how much the feature contributes to the prediction of the target variable. The redundancy of a feature is assessed using correlation scores, which indicate how much the feature overlaps with other selected features.This method considers both the F-statistic score and the correlation score to evaluate the relevance and redundancy of each feature. The mRMR score is calculated iteratively for each feature, considering its relevance with the target variable and its redundancy with previously selected features.

The mRMR (minimum redundancy and maximum relevance) with relevance and redundancy, can be calculated using the following equation:

$$\text{mRMR}(X_i) = \frac{\text{Relevance}(X_i|Y)}{\text{Redundancy}(X_i|\{X\}_{\text{features selected until } i-1})}$$

Where:

- $\text{Relevance}(X_i|Y)$ is the relevance of feature X_i with respect to the target variable Y.
- $\text{Redundancy}(X_i|\{X\}_{\text{features selected until } i-1})$ is the redundancy of feature X_i with respect to the features selected until iteration $i-1$.

The above equation calculates the ratio of relevance to redundancy for each feature X_i, helping to select features that are both relevant to the target variable and non-redundant with each other.The relevancy and non-redundant features selected are further optimized for selecting the most-minimal subset of features by measuring the joint mutual information score between the variables [26].The correlation measure in mRMR cannot measure the inter- relation between the variables and the different classes of target variables. On the contrary, Joint Mutual Information(JMI) score can capture this.The proposed feature selection method utilizes Joint Mutual Information (JMI) to quantify

the statistical interdependence between individual features and the target variable, aiming to capture their collective information. This study emphasizes the effectiveness of JMI in identifying significant features that exhibit strong correlations and interactions with the target variable. The calculation of the joint mutual information score involves determining the joint entropy of the feature variables and the target variable.JMI aims to select features that collectively provide the most relevant information about the target variable, enhancing interpretability and capturing joint information among the features. Features with JMI values greater than 0.5 are chosen and added to the feature subset.

The final feature set, F, is the set intersection of the features selected by mRMR and JMI:

$$F = \{X_i\}_{\text{mRMR}} \cap \{X_i\}_{\text{JMI}} \tag{1}$$

4.1.2 Algorithmic Weighted Approach for Class Weight Generation

The algorithmic weighted approach for imbalanced datasets is a technique used to address the class imbalance problem in machine learning. In imbalanced datasets, one class (the minority class) is significantly underrepresented compared to the other class (the majority class). This can lead to poor performance of machine learning models, as they tend to favor the majority class and may struggle to correctly classify instances from the minority class. The algorithmic weighted approach aims to balance the influence of each class in the model by assigning different weights to the classes during training. This means that errors made on the minority class are penalized more heavily than errors on the majority class. This approach allows the model to pay more attention to the minority class and improve its ability to correctly classify instances from this class.

4.1.3 Results and Analysis

The performance evaluation of mRMR_JMI&AW with the Random Forest Classifier was conducted using ROC analysis, as depicted in Figure 1 and 2. The Receiver Operating Characteristic (ROC) curve illustrates the relationship between the false positive rate and the false negative rate for the classifier under consideration, across different probability thresholds. The area under the ROC curve serves as a metric for evaluating the performance of the classifier.In Figure 1, it can be observed that the Area Under the Curve (AUC) score for all five classes is 0.9999, which suggests a remarkably high level of performance in effectively identifying between both the majority and minority classes. In the present study, the performance of the ToN-IoT model was evaluated with respect to its ability to classify various classes, namely Backdoor, Benign, DDoS, DoS, Password, Ransomware, Scanning, and XSS. As depicted in Figure 2, the area under the curve (AUC) scores for these classes were found to be within the range of 0.999 to 1.0. This observation

indicates the exceptional discriminatory capacity of the model in distinguishing between the aforementioned classes. Despite the limited sample size for MITM and Injection classes, the Receiver Operating Characteristic Area Under the Curve (ROC-AUC) scores exhibit a noteworthy range of 0.9679 to 0.9879.

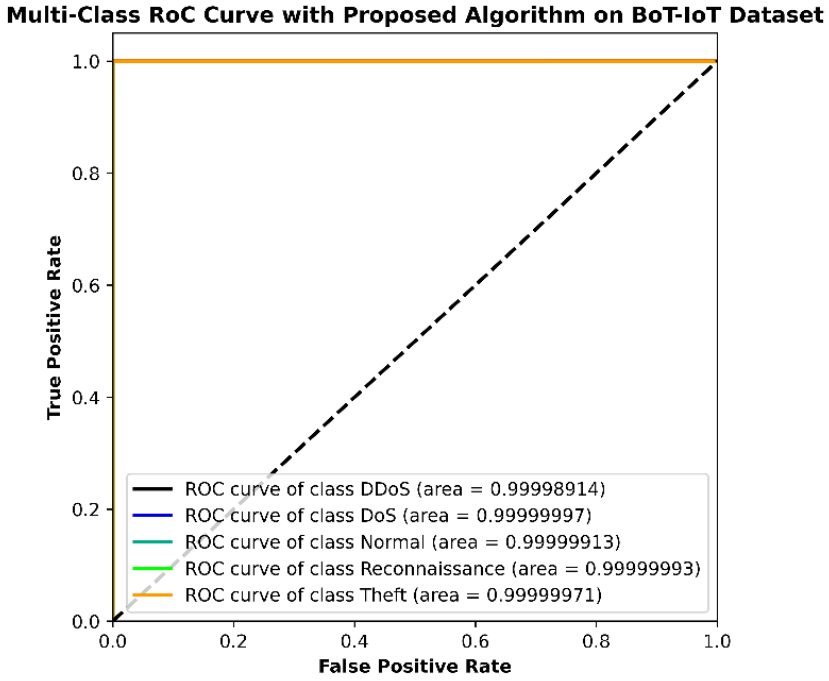

Multi-Class RoC Curve with Proposed Algorithm on BoT-IoT Dataset
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class DDoS (area = 0.99998914)
ROC curve of class DoS (area = 0.99999997)
ROC curve of class Normal (area = 0.99999913)
ROC curve of class Reconnaissance (area = 0.99999993)
ROC curve of class Theft (area = 0.99999971)

Figure 1: Multi-Class Roc Curve Analysis with BoT-IoT Dataset

4.2 Proposed Scheme for Detection of Known and Unknown Attacks

4.2.1 Proposed Scheme for Detection of Class-wise Feature Selection

In the proposed methodology, the feature selection process is designed to identify the most pertinent features that can effectively differentiate each class. The feature selection approach is based on a learning vector quantization (LVQ) method [27], which identifies prototypes that represent the various classes within the data. Each data sample is classified by calculating its distance from all prototypes and assigning it to the class of the nearest prototype. The NF-BoT-IoT dataset employed in the experimental study is denoted as $D = \{(x_i, y_i) \mid x_i \in \mathbb{R}^N, y_i \in \{1, \ldots, C\}\}_{i=1}^{P}$, where P represents the number of samples x_i with labels y_i across C classes. The LVQ model on D comprises several prototypes, identified by their positions in the weight space $w_i \in \mathbb{R}^N$ and their class labels $c(w_i) \in \{1, \ldots, C\}$. A data point x_i is classified using a similarity measure d^{λ}, where λ is the metric parameter that is adjustable during training. The data point x_i is mapped to

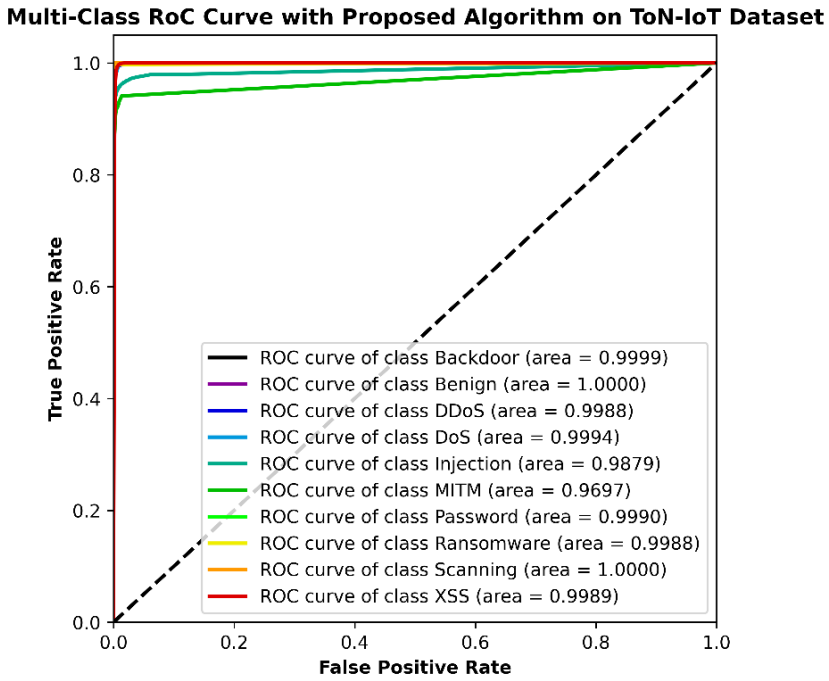

Multi-Class RoC Curve with Proposed Algorithm on ToN-IoT Dataset
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class Backdoor (area = 0.9999)
ROC curve of class Benign (area = 1.0000)
ROC curve of class DDoS (area = 0.9988)
ROC curve of class DoS (area = 0.9994)
ROC curve of class Injection (area = 0.9879)
ROC curve of class MITM (area = 0.9697)
ROC curve of class Password (area = 0.9990)
ROC curve of class Ransomware (area = 0.9988)
ROC curve of class Scanning (area = 1.0000)
ROC curve of class XSS (area = 0.9989)

Figure 2: Multi-Class RoC Curve Analysis with ToN-IoT Dataset

the label $c(x_i) = c(w_i)$ of the prototype i for which $d^{\lambda}(w_i, x_i) \leq d^{\lambda}(w_j, x_i)$, where $j \neq i$, thus assigning it to the class of the closest prototype.

The dataset used in this experimental study exhibits non-linear class boundaries. To achieve better class separation, a modified version of LVQ called Localized General Matrix Learning Vector Quantization (LGMLVQ) is employed [28], [29]. The LGMLVQ algorithm refines the prototype vectors using an update rule that integrates the standard LVQ update mechanism with a neighborhood function based on the distances between the training data and the prototype vectors. This approach emphasizes the selection of relevant features for each class by assessing their contributions to class separation through Fisher's discriminant ratio.

The within-class covariance matrix captures the variability of feature values within a specific class. It is determined by examining the differences between the feature values of samples within that same class. The within-class covariance matrix S_c for a class c is as follows:

$$S_c = \frac{1}{n_c - 1} \sum_{i=1}^{n_{\text{samples}}} (x_i - \mu_c)(x_i - \mu_c)^T$$

Where:

n_c is the number of samples in class c.
x_i represents the feature vector of the i-th sample in class c.
μ_c is the mean feature vector of all samples in class c.

On the other hand, the between-class covariance matrix measures the variability between different classes, capturing how the feature values of samples from various classes are distributed relative to each other. The between-class covariance matrix $S_{\neg c}$ for class c is calculated as follows:

$$S_{\neg c} = \sum_{i=1}^{C} n_i(\mu_i - \mu)(\mu_i - \mu)^T$$

Where:

C is the total number of classes.
n_i is the number of samples in class i.
μ_i is the mean feature vector of all samples in class i.
μ is the overall mean feature vector calculated across all samples and classes.

The trace of the within-class covariance matrix S_c multiplied by the inverse of the sum of S_c and $S_{\neg c}$is calculated as J_c.Similarly,the trace of the between-class covariance matrix $S_{\neg c}$ multiplied by the inverse of the sum of S_c and $S_{\neg c}$, is calculated in $J_{\neg c}$. The Fisher's discriminant ratio is calculated as $\frac{J_c}{J_{\neg c}}$.is a statistical measure used to evaluate the separation between two or more classes in a feature space. It quantifies how well the classes are separated while considering the within-class variability and the between-class variability of the feature values

Fisher's discriminant ratio is a statistical metric used to assess the degree of separation between classes within a feature space. The relevance matrix is constructed using Fisher's scores, and a threshold is applied as a criterion to select the most discriminative features. A Fisher's score exceeding 0.5 suggests that the feature contributes more to the variability between classes than to the variability within classes, ensuring that only those features

that significantly enhance class separation are chosen. The threshold of 0.5 is chosen to balance inclusiveness and selectivity, preventing the inclusion of less discriminative features while avoiding the exclusion of potentially valuable ones. This approach aids in creating a feature subset that captures key discriminative information, helping to filter out less informative or non-discriminatory features and thereby reducing the likelihood of overfitting.

4.2.2 Design of Stacked Multi-Classifier and Detection of Known Attacks

The pre-processed dataset is split into D' and D'' according to the 90:10 rule. D' is further split into five subdatasets D'_{Benign}, D'_{DDoS}, D'_{DoS}, $D'_{\text{Reconnaissance}}$, and D'_{Theft} based on the five classes. The classwise sub-datasets with only relevant features are split into training and testing sets according to the 80:20 rule.The training and testing classwise sub-datasets are used to build the proposed Multi-Layer Perceptron Classifier(MLP). $X^{\text{train}}_{\text{Benign}}$ and $Y^{\text{train}}_{\text{Benign}}$ are used to build the classifier $\text{MLP}_{\text{Benign}}$. The testing performance of the model is further evaluated using $X^{\text{test}}_{\text{Benign}}$ and $Y^{\text{test}}_{\text{Benign}}$. Similarly, MLP_{DDoS}, MLP_{DoS}, $\text{MLP}_{\text{Reconnaissance}}$, and $\text{MLP}_{\text{Theft}}$ classifiers are built and evaluated on the respective sub-datasets created for DDoS, DoS, Reconnaissance, and Theft classes. Furthermore, the D'' dataset, which has samples from all the classes, is used to evaluate the robustness in classifying the known attack classes. The batch samples consisting of all classes are fed to each individual MLP classifiers.The relavant features of Benign class are only selected while testing with the Benign MLP Classifer and the prediction score from each MLP_i for that batch is appended to prob$'$. Similarly the batch of network samples are evaluated using each of the stacked MLP classifier.The final prediction of the batch samples is obtained by choosing the prediction of the MLP classifier with the maximum prediction score.

4.2.3 Proposed Scheme for Detection of Unknown Attacks

The proposed design for the detection of unknown attacks involves two shallow autoencoders. The shallow autoencoder is a lightweight design comprising only one hidden layer or latent layer between the input and output layers. In the proposed approach, the benign samples of the BoT-IoT dataset is used to train the benign autoencoder represented as $\text{AE}_{\text{Benign}}$ and the attack samples from all the four classes DDoS, DoS, Reconnaissance, and Theft are combined to train the attack autoencoder $\text{AE}_{\text{Attack}}$. The $\text{AE}_{\text{Benign}}$ and $\text{AE}_{\text{Attack}}$ auto-encoders are comprised of an encoder and a decoder. The encoder utilizes a deterministic non-linear mapping function to map input x to represent the hidden layer, H. The encoding procedure can be written as $H = F(Wx + b)$ where W is the weight between the input and the hidden layer representation H, and b is the bias. The decoder reconstructs the hidden representation of the hidden layer to output $\hat{x}$ and $\hat{x}$ is represented

as $F'(W'H + b')$ where W' is the weight between the hidden layer representation and the output, b' is the deviation, and x' is the reconstruction output of Y.The shallow autoencoders $\text{AE}_{\text{Benign}}$ and $\text{AE}_{\text{Attack}}$ are trained by minimizing the reconstruction error. Consider the training data $D = \{x^i_{\text{Benign}}\}^N_{i=1}$ and $\{x^i_{\text{Attack}}\}^N_{i=1}$, the loss function of AE is $L(x, \hat{x}) = \frac{1}{N} \sum_{x \in D} \|x - \hat{x}\|^2$. The $\text{AE}_{\text{Benign}}$ trained with normal traffic data is capable of reconstructing any input as closely as feasible to the learned normal patterns. Similarly, the $\text{AE}_{\text{Attack}}$ trained with attack traffic will reconstruct any input as closely as possible to the learned attack patterns of the BoT-IoT dataset.

4.2.4 Result and Analysis

Performance Analysis of Proposed Known Attack Detection Methodology The proposed classifier was evaluated on a hold-out dataset, referred to as D'', which includes samples from all categories: 36,740 for Benign, 67,906 for DDoS, 48 for DoS, 123 for Reconnaissance, and 8 for Theft. Figure 8 illustrates the confusion matrix of the model after applying 10-fold cross-validation with the D'' NF-BoT-IoT dataset. The confusion matrix's diagonal elements show the number of correctly classified examples for each class. The matrix indicates that the false-positive rates for all classes are almost negligible, suggesting strong performance of the classifier. Specifically, the stacked classifier made no errors in the Benign, DDoS, DoS, and Theft categories, and only two errors in the Reconnaissance category. The confusion matrix further highlights the classifier's effectiveness in detecting various attack types, with all 8 Theft samples correctly classified without the need for resampling. Benign and DDoS classes have a significantly higher number of samples compared to DoS, Reconnaissance, and Theft.

The Stacked MLP classifiers, developed using the selected features for each class through the proposed LGLMVQ method, achieved accuracy, precision, recall, and F1 Score values ranging from 0.999 to 0.9989. The precision value of 0.9989 reflects an extremely low rate of false positive predictions. Furthermore, the model attained an outstanding RoC-AUC score of 0.9975, demonstrating its exceptional capability in effectively distinguishing and classifying both majority and minority classes. The ROC Curve and Precision- Recall curve of the detection model is drawn in Fig:4 and Fig:5. ROC curves depict the trade-off between TPR and FPR for a prediction model with various probability thresholds, whereas Precision–Recall curves describe the trade-off between the TPR and the positive predictive value for a predictive model with varying probability thresholds. As seen in Fig:4 and Fig:5, both curves distinguish all classes while covering 100 percent or close to 100 percent of the region. We evaluated the performance of our model individually for each class, and the micro-average value of our model's RoC and Precision–Recall curves indicates outstanding performance.The exceptionally high ROC scores for both minority and majority classes

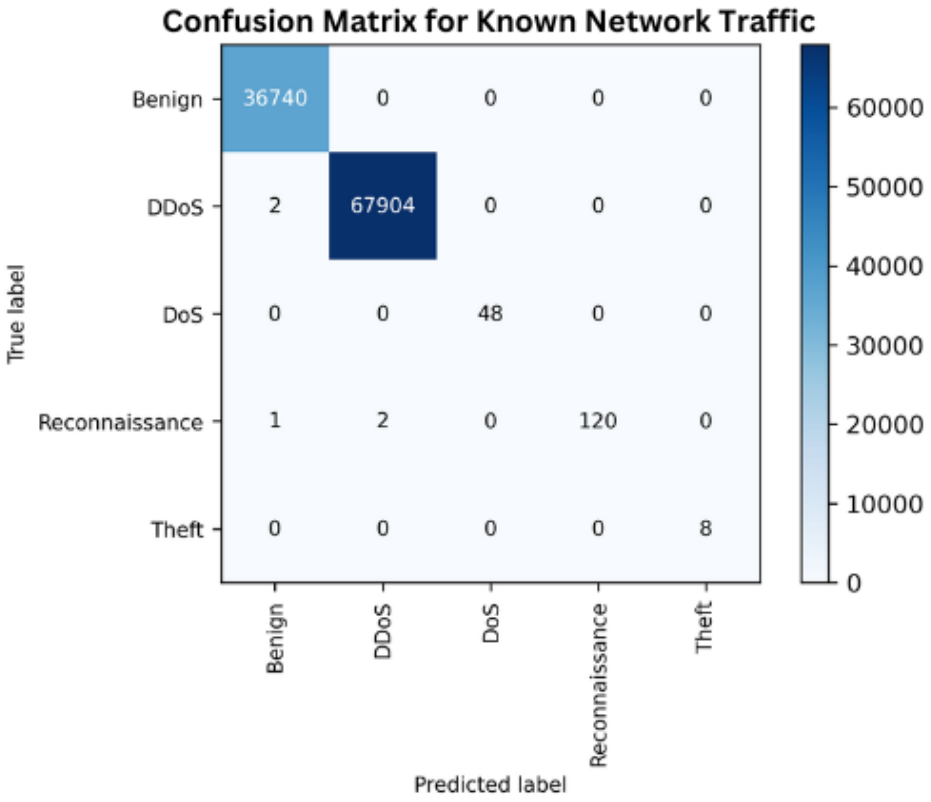


Figure 3: Confusion Matrix for Stacked Multi-Classifer for Known Attack Detection with BoT-IoT Dataset

suggest that the model has likely learned distinct and well-separated features for these classes, making their identification easier. The precision-recall curves are excellent metrics for evaluating the performance of the model under imbalanced scenario. The exceptionally high scores for all classes except Reconnaissance indicate very high precision and recall scores. The negligible misclassifications of the Reconnaissance class to Benign class have resulted in a score of 0.9756.The slightly lower PR curve for Reconnaissance" may be due to the more complex and subtle nature of reconnaissance activities, which can be challenging to distinguish from Benign traffic.

Performance Analysis of Proposed Unknown Attack Detection Methodology
The performance of the proposed two-level shallow autoencoder for detecting unknown attacks is evaluated using the ToN-IoT dataset. In this evaluation, labels were removed from 612,758 attack samples and 330,665 benign samples, and batches of unknown network samples were prepared. The method for distinguishing between benign and attack samples relies on setting specific benign and attack threshold values. After establishing these thresholds, any sample with a reconstruction error between the attack and benign thresholds is classified as an attack. Conversely, samples with reconstruction errors below the benign threshold but above the attack threshold are classified as benign. Samples

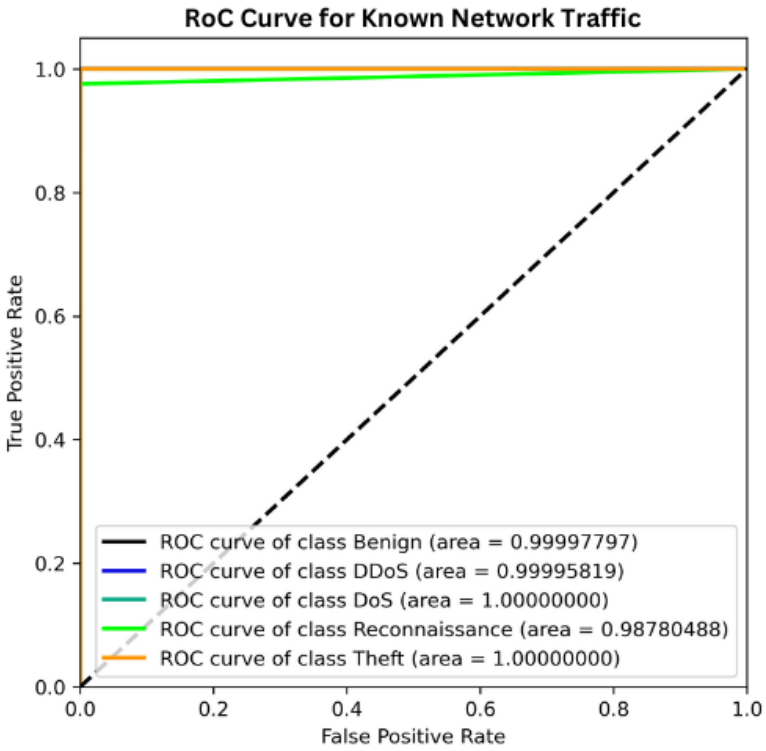

RoC Curve for Known Network Traffic
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve of class Benign (area = 0.99997797)
ROC curve of class DDoS (area = 0.99995819)
ROC curve of class DoS (area = 1.00000000)
ROC curve of class Reconnaissance (area = 0.98780488)
ROC curve of class Theft (area = 1.00000000)

Figure 4: RoC Curve for Stacked Multi-Classifier

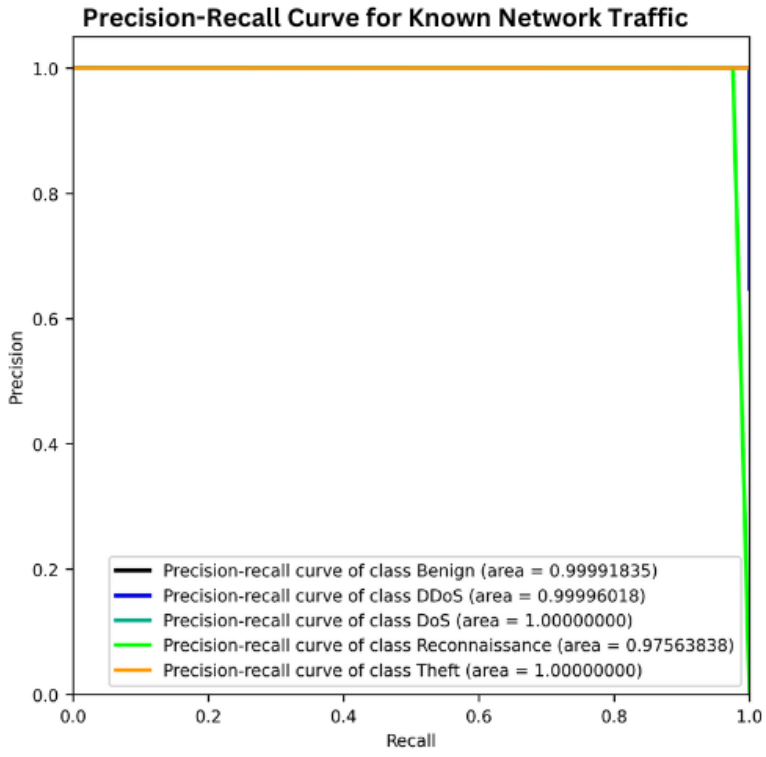


Figure 5: Precision-Recall Curve for Stacked Multi-Classifier

with reconstruction errors below both thresholds or above both thresholds are categorized as suspicious and require further analysis.

The benign autoencoder is trained on benign data, while the attack autoencoder is trained on attack samples from the BoT-IoT dataset. To assess the effectiveness of the proposed approach, a test set containing benign and attack samples from the ToN-IoT dataset is utilized. The results demonstrate that the method can accurately distinguish between unseen benign and attack samples, achieving high accuracy, precision, recall, and F1 scores. The misclassifications, as depicted in Figure 6, support the effectiveness of the proposed approach in detecting unknown samples. This indicates that the method is proficient in differentiating between benign, attack, and unknown samples, and shows promise for real-world application in detecting unknown attacks. In summary, the proposed method successfully detects unknown network samples by applying thresholds to the benign and attack autoencoders.

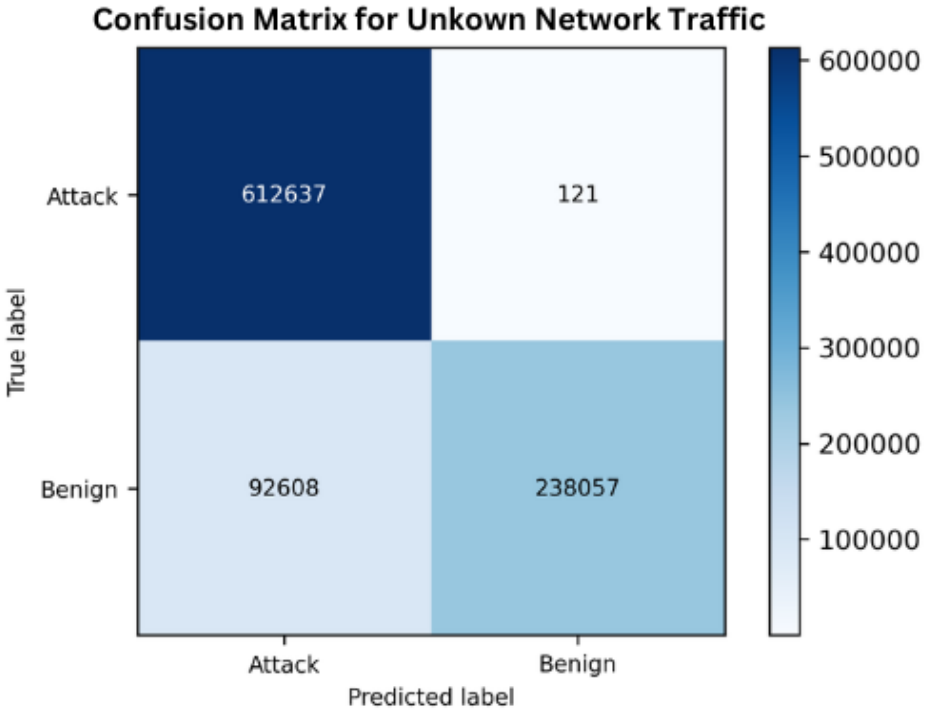


Figure 6: Confusion Matrix for Unknown Attack Detection with ToN-IoT Dataset

The generalizability of the proposed method for detecting unseen attacks is evaluated using the NF-CIC-IDS 2018 dataset, which includes samples of benign traffic as well as attack traffic from categories such as Bruteforce, Botnet, DoS, DDoS, Infiltration, and Web Attacks. The Benign and Attack Autoencoders, trained on the NF-BoT-IoT dataset, are assessed using unlabelled benign and attack traffic samples from NF-CIC-IDS 2018. The evaluation results, presented in the confusion matrix and classification report in Figure 7 and Table 2, indicate that only 14 out of 20,000 attack samples were misclassified.

Table 1: Performance Metrics on Unseen Network Samples

Dataset	Accuracy	Precision	Recall	F1 Score	False Positive Rate	False Negative Rate
NF-ToN-IoT	0.9017	0.9998	0.8687	0.9296	0.0005	0.1313
NF-CIC-IDS-2018	0.9726	0.9993	0.9486	0.9733	0.0007	0.0514

Table 2: Class-wise Performance Metrics for NF-CIC-IDS 2018

Class	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Attack	0.972	1.000	0.950	0.970
Benign	0.972	0.950	1.000	0.970

Similarly, 1,084 benign samples were incorrectly identified as attack traffic. These results are notable, as the method misclassifies very few attack samples as benign. The class-wise performance metrics in Table 2 demonstrate that the proposed technique effectively detects unseen attacks, achieving high accuracy, precision, recall, and F1 scores. Additionally, Table 1 highlights the average classification scores across different datasets, showing that the proposed method yields sufficiently high classification scores with minimal false positive and false negative rates.

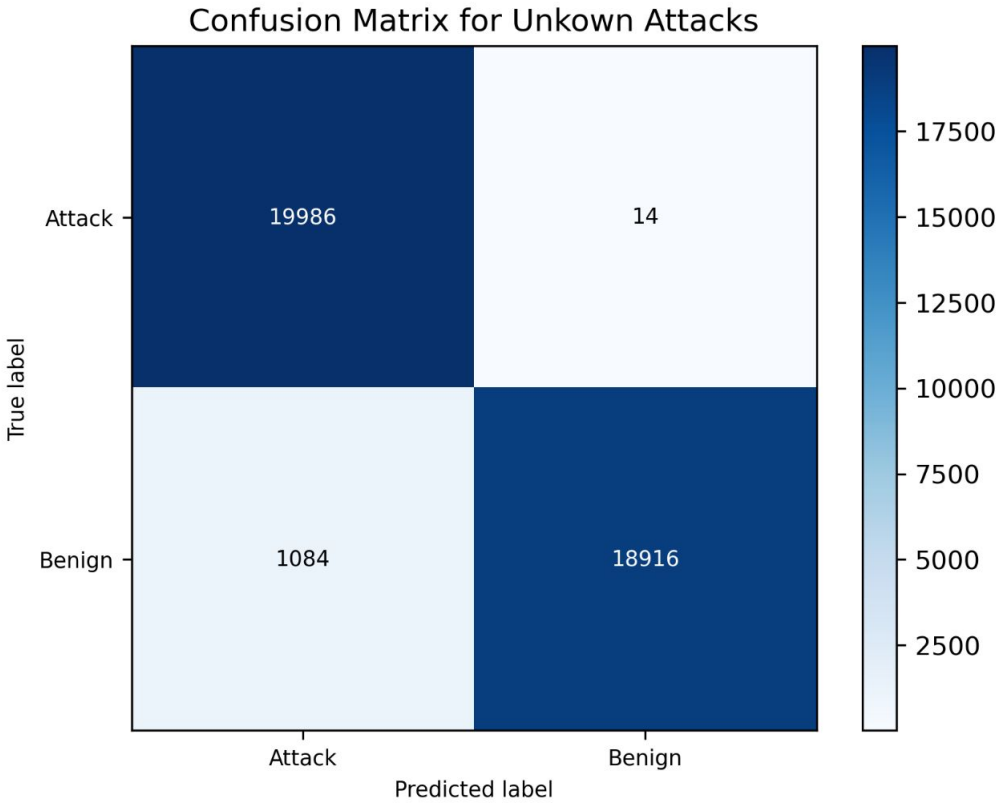


Figure 7: Confusion Matrix for Unknown Attack Detection with NF-CIC-IDS-2018

Several studies have explored unknown attack detection using similar datasets, with

Table 3: Class-wise Performance Metrics for NF-ToN-IoT

Class	Accuracy	Precision	Recall	F1 Score
Attack	0.9017	1.000	0.870	0.930
Benign	0.9017	0.720	1.000	0.840

one notable performance analysis of deep learning, supervised, and unsupervised machine learning classifiers conducted by [30]. This study used a dataset with attack samples distributed between 0.5% and 0.60% of the total, with the remaining samples consisting solely of known attacks. Among the supervised learning classifiers, the Decision Tree model demonstrated the highest performance, achieving an accuracy rate of 93.5% and a recall value of 73%. However, the recall for unidentified samples was found to be only 40%. In contrast, unsupervised learning techniques achieved a maximum recall of 58% for unknown samples. Meta-supervised and meta-unsupervised methods showed improved performance, with recall rates for unknown samples ranging from 38% to 65%.

The proposed method employing adaptive thresholding for detecting unknown attacks has demonstrated notably superior performance. Evaluation of this method on the NF-ToN-IoT dataset yields impressive results. The accuracy achieved is 0.9017, reflecting a high degree of correctness in identifying unknown attacks. Additionally, the precision of 0.9998 indicates very few false positives in the detection process. The recall, which assesses the method's ability to detect all attack instances, stands at 0.8687, showing a robust detection capability. The F1 score, which balances precision and recall, is 0.9296, suggesting a well-rounded performance in both metrics. Overall, these findings underscore the effectiveness of the proposed approach in accurately detecting unknown attacks within the NF-ToN-IoT dataset.

4.3 Design of Robust Botnet Detection Model against Adversarial Attacks

4.3.1 Proposed Defense using Shallow Denosing AutoEncoder

The proposed defense mechanism against adversarial attacks involves the utilization of a shallow Denoising Autoencoder (DAE). This DAE comprises an encoder and a decoder, working in tandem to compress and reconstruct data, respectively. The encoder processes the input dataset, which includes adversarial attack samples, and compresses it into a lower-dimensional representation. By minimizing the learning cost function, the DAE achieves dimensionality reduction, thereby reducing or eliminating adversarial noise. The DAE minimizes the loss function L in the following manner:

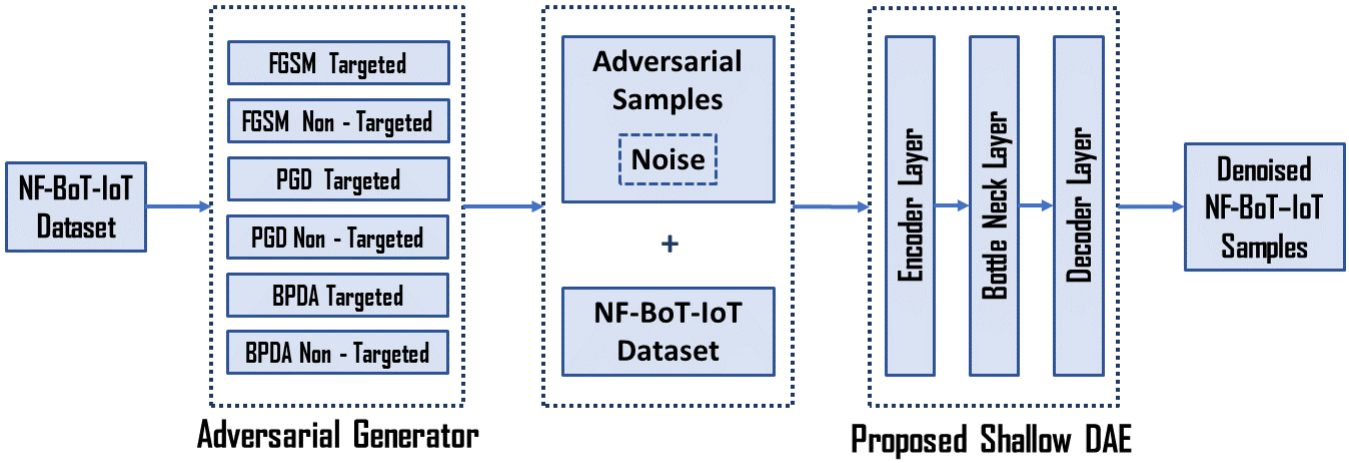


Figure 8: Illustration of Proposed Shallow DAE Defense Mechanism for Adversarial Attacks

$$L = \frac{1}{n}\sum_{i=1}^{n}(x_i - F_\theta(Wx_i + b))^2$$

where F represents nonlinear activation functions, W and b denote the encoder's weight and bias of the encoder. The decoder takes the compressed representation and reconstructs an approximation of the original input data. The DAE is trained to learn a mapping f that takes input data X of adversarial samples to a reconstructed output $\hat{X}$ with no adversarial samples. The training process minimizes the reconstruction error, measured using the loss function mean squared error (MSE):

$$\text{MSE Loss} = \frac{1}{N}\sum_{i=1}^{N}\|X_i - \hat{X}_i\|_2^2$$

where N is the number of training samples, X_i is the clean input, and $\hat{X}_i$ is the corresponding reconstructed output.

The design of the proposed DAE is illustrated in Fig.8. The DAE is trained to minimize the MSE between the original NF-BoT-IoT sample X and the DAE-generated output for a given input X. This indicates that during training, the DAE learns to reconstruct the original, clean input from the perturbed input, effectively denoising the data. The pre-processed NF-BoT-IoT dataset is split into training and testing in an 80:20 ratio. This dataset is pre-processed and then divided into training and testing sets using an 80:20 ratio. The entire NF-BoT-IoT dataset is used to generate FGSM Non-Targeted

Adversarial samples with a perturbation parameter of 0.25. The resulting adversarial samples are further split in an 80:20 ratio, with 80% of the samples, totaling 838,598, allocated for training the DAE. During training, 80% of the FGSM Non-Targeted samples form the input of the proposed DAE, and 80% of the pre-processed NF-BoT-IoT dataset with no noise will be the output. The denoising autoencoder is characterized by its lightweight architecture, comprising three layers: an input layer, a latent layer, and an output layer.

The hyperparameters used for designing the DAE are described in Table 4. The model is trained over 100 epochs using a batch size of 128. The design of the autoencoder is implemented using TensorFlow 2, with Keras serving as the backend framework. The optimization algorithm employed is Adam, with a learning rate set to 0.02. Furthermore, the rectified linear unit (ReLU) activation function is utilized for all fully connected layers, while the loss function is specified as the Mean Squared Error. The proposed shallow DAE is validated using 20% of the FGSM Non-Targeted samples as input. Experimental results indicate that after 100 epochs, the validation loss decreased to 0.0015, and the MSE decreased to 0.0025. The model achieved a state of stability, as there were no subsequent changes in the validation loss or MSE values. To further assess the effectiveness of the proposed DAE, we generated testing samples using various adversarial attack techniques, including FGSM Non-Targeted, FGSM Targeted, PGD Non-Targeted, and PGD Targeted. For each case, we considered 20% of the generated samples, resulting in 209,650 samples for FGSM Non-Targeted, FGSM Targeted, PGD Non-Targeted, and PGD Targeted attacks, respectively.

Table 4: Hyperparameters of Proposed Shallow Denoising Auto-Encoder

Model Parameters	**Values**
Input Neurons	39
Hidden Layer	-
Number of Neurons in Hidden Layer	-
Latent Layer	1
Number of Neurons in Latent Layer	10
Output Neurons	39
Optimizer Function	Adam
Activation Function	ReLU
Dropout	-
Loss Function	MSLE
Number of Epochs	100
Batch Size	128

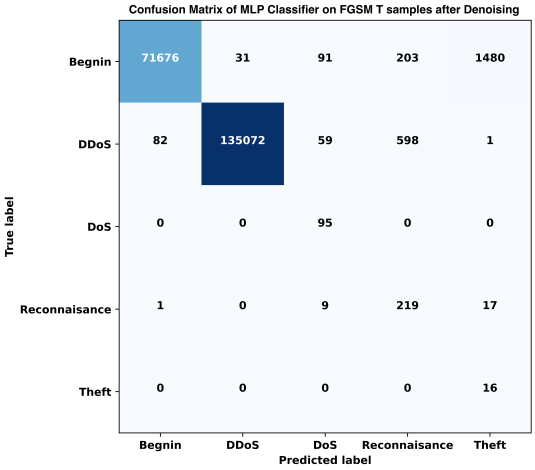
4.3.2 Result and Analysis

In this section, we conducted a comparative analysis between the proposed defense mechanism employing Shallow Denoising Autoencoder and the existing Adversarial Training approach. Our evaluation focused on comparing the effectiveness of these defenses by assessing the classification performance of the weighted MLP model against various categories of adversarial attacks. The Denoising Autoencoder is trained using FGSM Non-Targeted samples and subsequently tested with adversarial attack samples from each category. Similarly, in the case of adversarial training, the weighted MLP model is trained using FGSM Non-Targeted samples and then tested with adversarial samples from all four adversarial attack categories. We investigated how the classification scores of the MLP model differed when subjected to adversarial samples, comparing the proposed Denoising Autoencoder defense with Adversarial Training. We further elaborated on the differences observed in the confusion matrix and RoC Curve of the MLP model under each category of attacks, highlighting the impact of each defense mechanism.

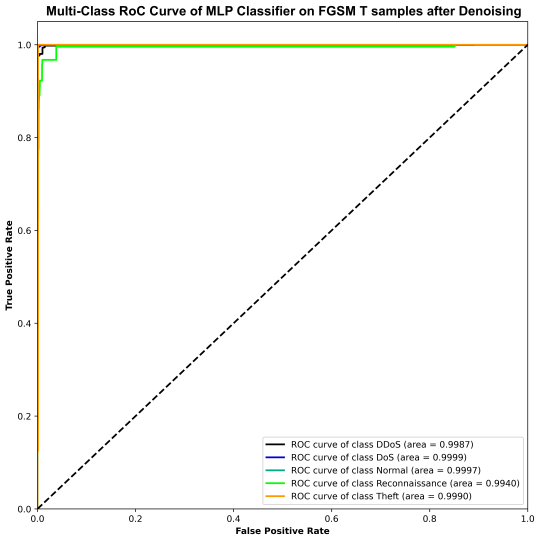
Targeted FGSM Attack Upon examination of the results using the confusion matrix in Fig.9, it is evident that the Denoising Autoencoder method consistently outperformed the adversarial training approach across various network traffic classes in defending against FGSM targeted adversarial attacks. Specifically, the Denoising Autoencoder method achieved superior performance in correctly classifying benign, DoS, and theft samples, while maintaining minimum misclassifications in the DDoS and reconnaissance classes. For instance, the number of misclassifications for the benign class when using the proposed Denoising Autoencoder was 1805 out of 73481 benign samples. However, in the case of Adversarial Training, all 73481 benign samples were misclassified.

Moreover, the proposed defense yielded zero misclassifications for the DoS and theft categories, which are the minority classes. In contrast, Adversarial Training misclassified all samples of the DoS, reconnaissance, and theft categories. The ROC-AUC curve analysis in Fig. 9 further demonstrated the robustness of the Denoising Autoencoder in maintaining the classification ability of the Weighted MLP model during FGSM Targeted Attacks. Conversely, the adversarial training approach exhibited substantial limitations in maintaining the classification power of the Weighted MLP Model, as indicated by the significantly higher number of misclassifications across all classes and lower RoC-AuC Score.

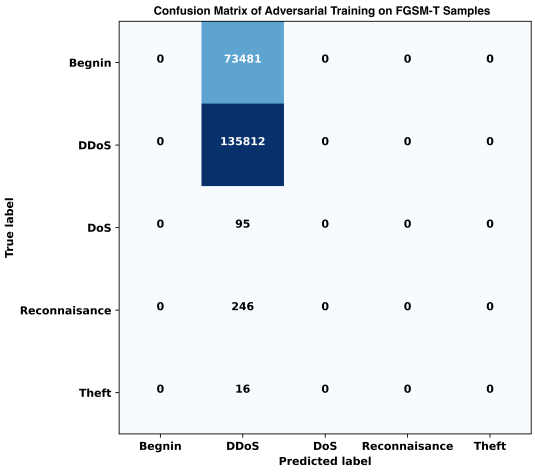
Non-Targeted FGSM Attack In the case of the proposed defense against FGSM Non-Targeted adversarial attacks, the confusion matrix depicted in Fig.10 revealed some misclassifications primarily in the Benign and DDoS categories, with very few misclassifi-



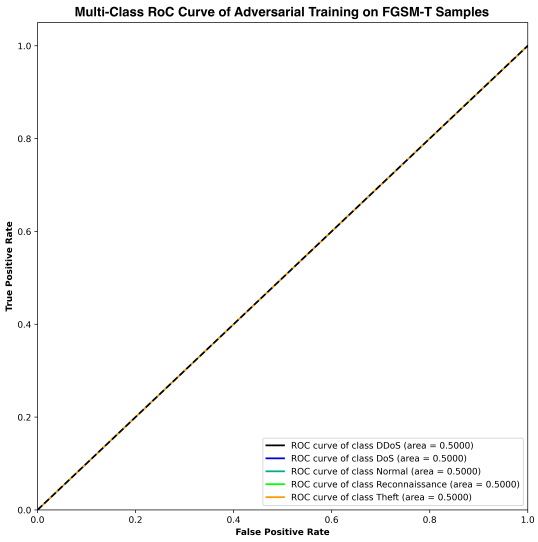
(a) Confusion Matrix with Proposed Defense


Multi-Class RoC Curve of MLP Classifier on FGSM T samples after Denoising
True Positive Rate
False Positive Rate
ROC curve of class DDoS (area = 0.9987)
ROC curve of class DoS (area = 0.9999)
ROC curve of class Normal (area = 0.9997)
ROC curve of class Reconnaissance (area = 0.9940)
ROC curve of class Theft (area = 0.9990)

(b) RoC Curve with Proposed Defense



(c) Confusion Matrix with Adversarial Training

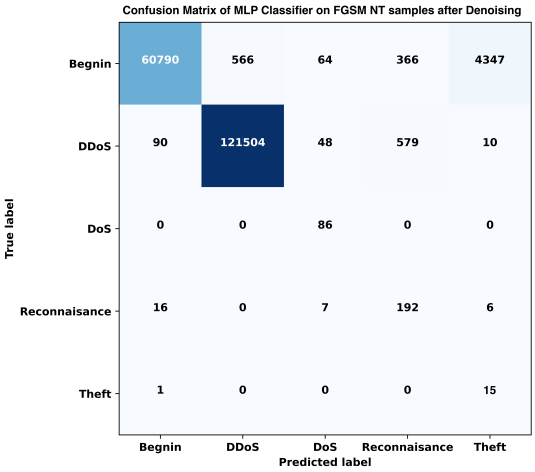

Multi-Class RoC Curve of Adversarial Training on FGSM-T Samples
True Positive Rate
False Positive Rate
ROC curve of class DDoS (area = 0.5000)
ROC curve of class DoS (area = 0.5000)
ROC curve of class Normal (area = 0.5000)
ROC curve of class Reconnaissance (area = 0.5000)
ROC curve of class Theft (area = 0.5000)

(d) RoC Curve with Adversarial Training

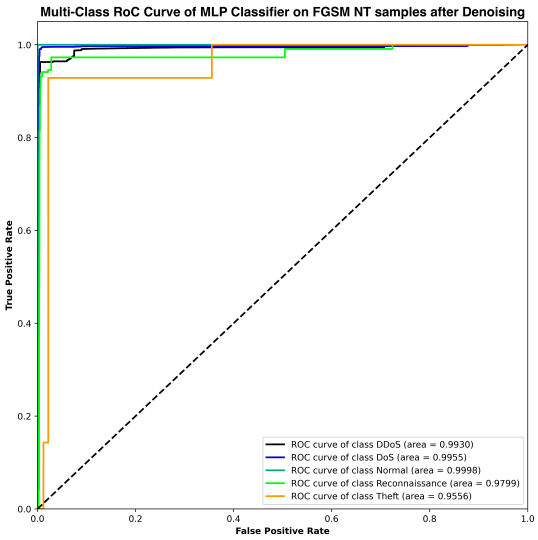
Figure 9: Comparative Analysis of Proposed Defense and Adversarial Training Approach under Targeted FGSM Attacks

cations in the Reconnaissance and Theft categories. In contrast, the Adversarial Training approach resulted in fewer misclassifications for the Benign and DDoS classes, but all samples in the DoS, Reconnaissance, and Theft categories were misclassified.

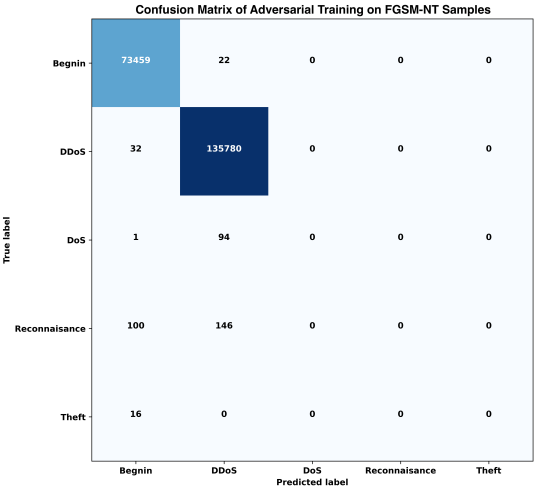
These findings indicate that the misclassification rates were consistently higher when employing adversarial training compared to our proposed Denoising Autoencoder approach. This observation is further supported by the ROC-AUC scores, which confirmed the outstanding classification capabilities of the weighted MLP when using the proposed defense. Conversely, the adversarial training approach resulted in lower classification scores for the weighted MLP model under FGSM Non-Targeted Attacks, indicating the inability to defend FGSM Non-Targeted attacks.



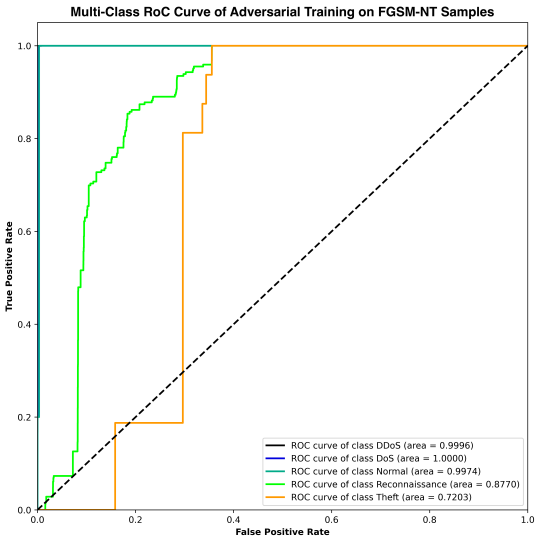
(a) Confusion Matrix with Proposed Defense


Multi-Class RoC Curve of MLP Classifier on FGSM NT samples after Denoising
True Positive Rate
False Positive Rate
ROC curve of class DDoS (area = 0.9930)
ROC curve of class DoS (area = 0.9955)
ROC curve of class Normal (area = 0.9998)
ROC curve of class Reconnaissance (area = 0.9799)
ROC curve of class Theft (area = 0.9556)

(b) RoC Curve with Proposed Defense



(c) Confusion Matrix with Adversarial Training


Multi-Class RoC Curve of Adversarial Training on FGSM-NT Samples
True Positive Rate
False Positive Rate
ROC curve of class DDoS (area = 0.9996)
ROC curve of class DoS (area = 1.0000)
ROC curve of class Normal (area = 0.9974)
ROC curve of class Reconnaissance (area = 0.8770)
ROC curve of class Theft (area = 0.7203)

(d) RoC Curve with Adversarial Training

Figure 10: Comparitive Analysis of Proposed Defense with Adversarial Training Approach under Non-Targeted FGSM Attacks

The Denoising Autoencoder method effectively preserved the classification ability of the weighted MLP model, particularly in correctly classifying Benign, DoS, and Theft samples, while maintaining minimum misclassifications in the DDoS and Reconnaissance classes. In contrast, the adversarial training approach showed limitations in maintaining classification power, as indicated by higher misclassification rates and lower RoC-AUC scores across all classes. These findings underscore the effectiveness of the Denoising Autoencoder as a defense mechanism for enhancing the robustness of machine learning models against adversarial attacks.

5 Conclusion

This research addresses critical challenges in botnet detection within IoT networks, focusing on class imbalance and adversarial robustness. The following conclusions summarize the significant contributions and findings of this work:

1. **Multi-Stage Feature Selection and Algorithmic Weighted Approach:**
    - Developed a computationally efficient classification method for managing imbalanced datasets.
    - Implemented a multi-stage feature selection process integrating joint mutual information scores with maximum relevance and minimum redundancy.
    - Combined with an algorithmic weighted mechanism, this approach significantly improved classification performance without the need for re-sampling.
2. **Class-Wise Feature Selection and Adaptive Reconstruction Error Thresholding:**
    - Introduced a stacked Multilayer Perceptron (MLP) classifier optimized with class-specific features selected through the Localized Generalized Matrix Learning Vector Quantization (LGMLVQ) algorithm.
    - Demonstrated exceptional performance in detecting both known and unknown attacks.
    - Utilized shallow autoencoders with adaptive reconstruction error thresholding to effectively identify zero-day attacks.
3. **Adversarial Attack Generation and Resilient Model Design:**
    - Validated the efficacy of a shallow Denoising Autoencoder as a defense mechanism against adversarial attacks.
    - Outperformed traditional adversarial training methods, maintaining high classification accuracy and lower misclassification rates under various adversarial scenarios.

In summary, the methodologies developed and validated in this research offer substantial improvements in the detection and classification of botnet attacks in IoT networks. By addressing key issues such as class imbalance and adversarial threats, the proposed techniques contribute to more effective and reliable security solutions in the evolving landscape of IoT environments.

6 Future Scope

In the realm of enhanced adversarial robustness, future research could delve into exploring more sophisticated adversarial attack strategies beyond the commonly studied ones. This exploration could include black-box attacks, where the attacker has no knowledge of the model's internal workings, and transfer attacks, where adversarial examples crafted for one model are used to attack another. To counter such advanced attacks, robust defense mechanisms could be developed, drawing inspiration from the field of adversarial machine learning. Techniques like defensive distillation, which involves training the model on softened probabilities rather than raw logits, and adversarial training, which augments the training data with adversarial examples, could be integrated into the botnet detection framework to bolster its resilience against adversarial attacks.

In the context of adapting to concept drift, future research could focus on developing and enhancing online learning algorithms that can adapt to real-time changes in data distribution. These algorithms would enable the botnet detection framework to maintain its effectiveness and accuracy even as the characteristics of the data evolve over time. By continuously monitoring and adjusting to these changes, the model can remain robust and reliable in dynamic IoT environments, where concept drift is a common occurrence. Additionally, techniques such as ensemble learning, where multiple models are trained and combined to improve performance, could be explored to enhance the model's adaptability to concept drift.

7 Publications

Journal

- Krishnan, Deepa, Shrinath Pravin, "Robust Botnet Detection Approach for Known and Unknown Attacks in IoT Networks Using Stacked Multi-classifier and Adaptive Thresholding", *Arabian Journal for Science and Engineering*, Springer Publishers (Scopus, SCIE and WoS Indexed I.F 2.6) Available at `https://link.springer.com/article/10.1007/s13369-024-08742-y`

- Krishnan, Deepa, Shrinath Pravin, "Robust IoT Botnet Detection Framework Resilient to Gradient Based Adversarial Attacks", *SN Computer Science*, Springer Publishers (Scopus, ACM Digital Library)

Conferences

- Krishnan, Deepa, and Preeja Babu. "Imbalanced classification for botnet detection in Internet of Things." *Next Generation of Internet of Things: Proceedings of ICNGIoT 2021*, GIET Univerity, India, Springer, 2021.

- Krishnan, Deepa, and Preeja Babu. "An Adaptive weighted voting classifier for botnet detection in internet of things." *2021 IEEE 18th India Council International Conference (INDICON)*. IIT Guwahati, IEEE, 2021.

References

[1] Gaurav Sharma, Stilianos Vidalis, Niharika Anand, Catherine Menon, and Somesh Kumar. A survey on layer-wise security attacks in iot: Attacks, countermeasures, and open-issues. *Electronics*, 10(19):2365, 2021.

[2] Ismail Butun, Patrik Österberg, and Houbing Song. Security of the internet of things: Vulnerabilities, attacks, and countermeasures. *IEEE Communications Surveys & Tutorials*, 22(1):616–644, 2019.

[3] Ihsan Ali, Abdelmuttlib Ibrahim Abdalla Ahmed, Ahmad Almogren, Muhammad Ahsan Raza, Syed Attique Shah, Anwar Khan, and Abdullah Gani. Systematic literature review on iot-based botnet attack. *IEEE access*, 8:212220–212232, 2020.

[4] Nazrul Hoque, Dhruba K Bhattacharyya, and Jugal K Kalita. Botnet in ddos attacks: trends and challenges. *IEEE Communications Surveys & Tutorials*, 17(4):2242–2270, 2015.

[5] Simon Nam Thanh Vu, Mads Stege, Peter Issam El-Habr, Jesper Bang, and Nicola Dragoni. A survey on botnets: Incentives, evolution, detection and current trends. *Future Internet*, 13(8):198, 2021.

[6] Seungjin Lee, Azween Abdullah, Nz Jhanjhi, and Sh Kok. Classification of botnet attacks in iot smart factory using honeypot combined with machine learning. *PeerJ Computer Science*, 7:e350, 2021.

[7] Giuseppe Nebbione and Maria Carla Calzarossa. Security of iot application layer protocols: Challenges and findings. *Future Internet*, 12(3):55, 2020.

[8] Osama AlKadi, Nour Moustafa, Benjamin Turnbull, and Kim-Kwang Raymond Choo. Mixture localization-based outliers models for securing data migration in cloud centers. *IEEE Access*, 7:114607–114618, 2019.

[9] Long-Hao Yang, Tian-Yu Ren, Fei-Fei Ye, Peter Nicholl, Ying-Ming Wang, and Haitian Lu. An ensemble extended belief rule base decision model for imbalanced classification problems. *Knowledge-Based Systems*, 242:108410, 2022.

[10] Vivek Christopher, Tharmasanthiran Aathman, Kayathiri Mahendrakumaran, Rashmika Nawaratne, Daswin De Silva, Vishaka Nanayakkara, and Damminda Alahakoon. Minority resampling boosted unsupervised learning with hyperdimensional computing for threat detection at the edge of internet of things. *IEEE Access*, 9:126646–126657, 2021.

[11] Prabhat Kumar, Randhir Kumar, Gautam Srivastava, Govind P Gupta, Rakesh Tripathi, Thippa Reddy Gadekallu, and Neal N Xiong. Ppsf: A privacy-preserving and secure framework using blockchain-based machine-learning for iot-driven smart cities. *IEEE Transactions on Network Science and Engineering*, 8(3):2326–2341, 2021.

[12] Giang L Nguyen, Braulio Dumba, Quoc-Dung Ngo, Hai-Viet Le, and Tu N Nguyen. A collaborative approach to early detection of iot botnet. *Computers & Electrical Engineering*, 97:107525, 2022.

[13] Shahbaz Ahmad Khanday, Hoor Fatima, and Nitin Rakesh. Implementation of intrusion detection model for ddos attacks in lightweight iot networks. *Expert Systems with Applications*, 215:119330, 2023.

[14] Abdallah R Gad, Ahmed A Nashat, and Tamer M Barkat. Intrusion detection system using machine learning for vehicular ad hoc networks based on ton-iot dataset. *IEEE Access*, 9:142206–142217, 2021.

[15] Vikash Kumar and Ditipriya Sinha. Synthetic attack data generation model applying generative adversarial network for intrusion detection. *Computers & Security*, 125:103054, 2023.

[16] Laurens D'hooge, Tim Wauters, Bruno Volckaert, and Filip De Turck. Inter-dataset generalization strength of supervised machine learning methods for intrusion detection. *Journal of Information Security and Applications*, 54:102564, 2020.

[17] Rasheed Ahmad, Izzat Alsmadi, Wasim Alhamdani, and Lo'ai Tawalbeh. A deep learning ensemble approach to detecting unknown network attacks. *Journal of Information Security and Applications*, 67:103196, 2022.

[18] Nour Moustafa, Marwa Keshk, Kim-Kwang Raymond Choo, Timothy Lynar, Seyit Camtepe, and Monica Whitty. Dad: A distributed anomaly detection system using ensemble one-class statistical learning in edge networks. *Future Generation Computer Systems*, 118:240–251, 2021.

[19] Eva Rodríguez, Pol Valls, Beatriz Otero, Juan José Costa, Javier Verdú, Manuel Alejandro Pajuelo, and Ramon Canal. Transfer-learning-based intrusion detection framework in iot networks. *Sensors*, 22(15):5621, 2022.

[20] Zhangying He, Amin Rezaei, Houman Homayoun, and Hossein Sayadi. Deep neural network and transfer learning for accurate hardware-based zero-day malware detection. In *Proceedings of the Great Lakes Symposium on VLSI 2022*, pages 27–32, 2022.

[21] Rasheed Ahmad, Izzat Alsmadi, Wasim Alhamdani, and Lo'ai Tawalbeh. Zero-day attack detection: a systematic literature review. *Artificial Intelligence Review*, 56(10):10733–10811, 2023.

[22] Andrew McCarthy, Essam Ghadafi, Panagiotis Andriotis, and Phil Legg. Defending against adversarial machine learning attacks using hierarchical learning: A case study on network traffic attack classification. *Journal of Information Security and Applications*, 72:103398, 2023.

[23] Marek Pawlicki, Michał Choraś, and Rafał Kozik. Defending network intrusion detection systems against adversarial evasion attacks. *Future Generation Computer Systems*, 110:148–154, 2020.

[24] Giovanni Apruzzese, Mauro Andreolini, Mirco Marchetti, Vincenzo Giuseppe Colacino, and Giacomo Russo. Appcon: Mitigating evasion attacks to ml cyber detectors. *Symmetry*, 12(4):653, 2020.

[25] Hongling Jiang, Jinzhi Lin, and Haiyan Kang. Fgmd: A robust detector against adversarial attacks in the iot network. *Future Generation Computer Systems*, 132:194–210, 2022.

[26] Mohamed Bennasar, Yulia Hicks, and Rossitza Setchi. Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42(22):8520–8532, 2015.

[27] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[28] Atsushi Sato and Keiji Yamada. Generalized learning vector quantization. *Advances in neural information processing systems*, 8, 1995.

[29] Petra Schneider, Michael Biehl, and Barbara Hammer. Adaptive relevance matrices in learning vector quantization. *Neural computation*, 21(12):3532–3561, 2009.

[30] Tommaso Zoppi, Andrea Ceccarelli, Tommaso Puccetti, and Andrea Bondavalli. Which algorithm can detect unknown attacks? comparison of supervised, unsupervised and meta-learning algorithms for intrusion detection. *Computers & Security*, page 103107, 2023.